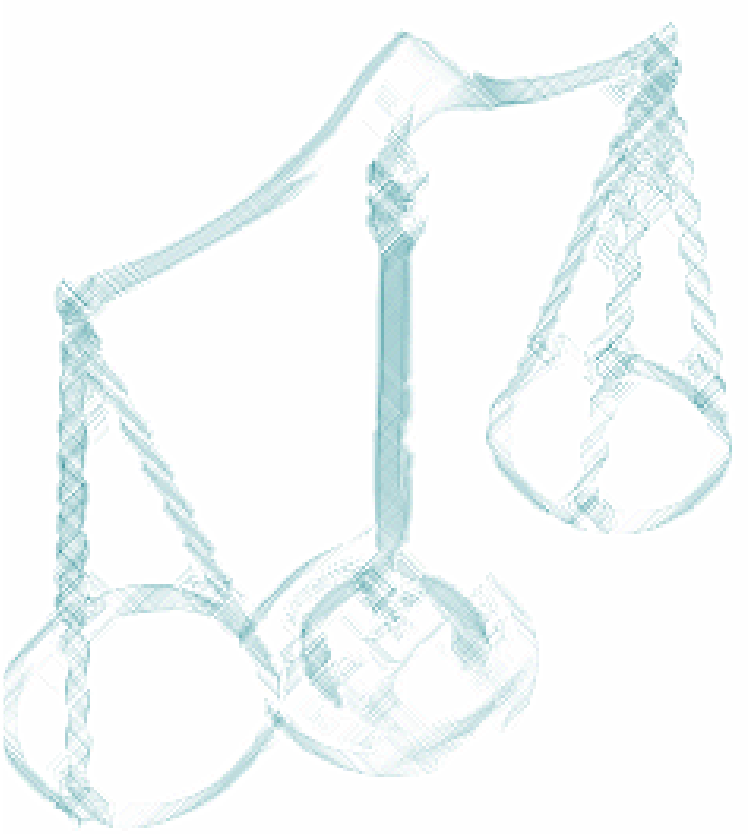
Institute for Energy

Formal Expert Judgement
An Overview



Authors: *K. Simola, A. Mengolini, R. Bolado-Lavin*

EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

EUR 21772 EN

Mission of the Institute for Energy

The Institute for Energy provides scientific and technical support for the conception, development, implementation and monitoring of community policies related to energy.
Special emphasis is given to the security of energy supply and to sustainable and safe energy production.

European Commission
Directorate-General Joint Research Centre (DG JRC)
Institute for Energy
Petten
The Netherlands

Contact:
Anna Mengolin Tel.: +31 (0) 224 56 5253
E-mail: anna.mengolini@jrc.nl

http://ie.jrc.cec.eu.int/

http://wwwjrc.cec.eu.int/

Luxembourg: Office for Official Publications of the European Communities, 2005

EUR 21772 EN

Printed in the Netherlands,(DG JRC, Institute for Energy , PR & Communication)

Cover: R. Houghton, JRC IE, PR & Communication
(No commercial use. Credit "Audiovisual Library European Commission".)

FORMAL EXPERT JUDGEMENT

AN OVERVIEW

K. Simola, A. Mengolini & R. Bolado-Lavin

July 2005

TABLE OF CONTENTS

FOREWORD

This document has been produced at JRC, Institute for Energy, within the frame of the institutional action SAFELIFE - Safety of Ageing Components in Nuclear Power Plants. SAFELIFE provides an integrated approach to R&D activities on critical issues for plant life management on ageing nuclear power installations.

The scope of the document is to provide an overview on the formal process of expert judgement and it is targeted to readers that are not very familiar with the issue, but are interested to have a condensed summary on the topic. An emphasis is put on the use of formal expert judgement in the field of structural integrity, this being an area of interest in plant life management of ageing nuclear power installations. Moreover, the document can also be seen as a basis for developing an approach for formal expert judgement in other plant life management areas such as, for example, maintenance and in-service inspection.

Although formal expert judgement has become a relatively well-established tool in connection to risk assessments, its application in the field of structural integrity of nuclear power plant components seems to be rather limited. However, we can see a growing need for moving towards probabilistic modelling in this field, one good example being the risk-informed in-service inspection (RI-ISI) methodology, where failure probabilities of piping or other structural components have to be evaluated. The lack of validated structural reliability tools and the scarcity of operating experience well justifies the consideration for the use of structured expert judgement.

1 INTRODUCTION

Judgements are inferences or evaluations that go beyond objective statements of fact, data, or the convention of a discipline. Judgements can be based on “facts” (factual judgements) that is to say on propositions that can be proven to be right or wrong, or they can be based on values (value judgements) that is to say on preferences among alternatives supported by priorities or tradeoffs. Many judgements require a special expertise, such as expertise in policy and decision-making, data interpretation, statistics, technical system, etc. Therefore a judgement that requires a special expertise is defined as “expert judgement”.

Expert judgement has been used in different fields to solve problems that can vary from very simple to very complicated and as an important input for decision-making. In many critical infrastructures of society, the parameters necessary for modelling physical, chemical or, sometimes, biological behaviour are not known with certainty. Therefore, when the value of an uncertain quantity is needed, and limits in data or understanding preclude the use of conventional statistical techniques to produce probabilistic estimates, the only option is to ask experts for their best professional judgements. Experts may indeed have valuable knowledge about models and parameters for problems in their specific field of interest.

Expert judgement can be the result of informal or formal processes. It is necessary to distinguish between formal expert judgement and informal expert judgement processes, the latter being the way expert judgement has traditionally been used. Expert judgement has been used in analysis and assessments in informal ways, through the expert’s implicit and undocumented reasoning, inferences, and scientific knowledge. In contrast, more recent formal uses of expert judgement exist that are explicit, structured and well documented. They try to bring out assumptions and reasoning that are at the base of a judgement, to quantify and document it so that others can appraise it.

Although formal judgement can be resource consuming (in terms of cost and time) and appear less flexible and creative than informal judgement, it brings out a deeper understanding of the issues, drawing out diverse opinions. This can reveal areas of disagreement or agreement with an increased possibility to seek solution. Furthermore, formal judgement gives access to all aspects of the analysis (assumptions, models, data, expert’s thinking) and provides the possibility for others to review the process of judgement.

Expert judgement, as mentioned earlier, is an important input in decision-making. Since decision analysis requires assumptions and reasoning underlying the judgement to be explicit, structured and well documented, it should rely on a formal process for extracting and quantifying the subjective uncertainty of experts. A formal process is also advisable because an accurate subjective probability judgement cannot be obtained by simply asking someone to provide a

probability number. Experts may not be familiar with expressing their uncertainties as probabilities. Further, the methods of reasoning used by an individual when generating a probability over a defined situation introduce biases into the number produced. Therefore, a formal and well-structured process may help in avoiding these pitfalls.

The present report focuses attention on formal expert judgement. It provides a brief description on the history of expert judgement (chapter 2), then focusing on the approach to formal expert judgement (chapter 3). Biases introduced in the process of elicitation (the process of extracting and quantifying the subjective judgements about uncertain quantities) by individual experts' way of reasoning are then presented (chapter 4) together with approaches on how these biases can be dealt with and the elicitation process aided through training (chapter 5). Then, methods of aggregation of expert's judgement are presented (chapter 6). Finally, an overview on recent and less recent application of expert judgement is given, with special attention to application in Structural Reliability.

2 HISTORY OF EXPERT JUDGEMENT

Formal expert judgement can be dated back to the 50's, when the USA experienced growth and faith in the use of expert opinion. In particular, the use of expert judgement can be linked to the establishment of the RAND (an acronym for Research and Development) Corporation after the World War II. World War II had revealed the significance of technology research and development for success in the battlefield and the importance of the wide range of scientists who made such development possible. A need was felt for an organization to connect military planning with research and development decisions. RAND Corporation was therefore founded to provide objective analysis and effective solutions to decision makers. Throughout the 1950s and beginning of 1960s, RAND Corporation worked almost exclusively for the Air Force. Later, it diversified its range of activities in the attempt to "demilitarise" its public image and attract new clients through the identification of other fields for application (Cooke, 1991).

Between the 1950s and the 1970s, the form in which structured expert opinion was conveyed was dictated by two methodologies both developed by RAND Corporation: **Scenario Analysis** and the **Delphi Method** which was then extended into **Cross Impact Analysis**. In the following we shortly summarise these three methodologies. For further references, see e.g. Cooke (1991).

2.1 The Scenario Analysis

In the definition of Hermann Kahn (Kahn, 1960), considered as the father of the scenario analysis, scenarios are "*hypothetical sequences of events constructed for the purpose of focusing attention on causal processes and decision points*" and they answer to two kinds of questions: 1) how might some hypothetical situation come about, step by step and 2) what alternatives exist for each stakeholders, at each step, to prevent, divert or facilitate the process.

The method can be summarized as follows:

1) The analyst identifies what he/she thinks is a set of basic long-term trends.
2) These trends are then extrapolated considering any theoretical or empirical knowledge that may have impact on the extrapolations. The result of this step is the so-called *surprise-free scenario*.
3) Based on the surprise-free scenario, other alternative scenarios can be defined, varying key parameters in the surprise free scenario.

There are no probabilities linked to the various scenarios since, in doing long range projections, the problem is that there are no 'more likely scenarios' than others. The surprise-free scenario is relevant because it is related to the basic long-term trends, not because it is probable. However "being relevant" doesn't

say much about prediction, therefore one shouldn't believe that a scenario yields predictions.

In conclusion, scenario analysis and the surprise free scenario that it supplies does not provide any probabilities. Therefore, the only use of studying the surprise free scenario and its alternatives is to gain a better comprehension of the basic long-term trends.

2.2 The Delphi Method

The Delphi method was developed in the early 1950s. In the middle of 1960s and early 1970s it found a wide variety of applications (Dalkey, 1968 and Brown et al., 1969). The method was mainly applied to technology forecasting, but also to many types of policy analyses.

The method is based on a structured process for collecting and distilling knowledge from a group of experts by means of a series of questionnaires combined with controlled opinion feed back.

The process can be summarized as follows:

- A questionnaire is sent to experts.
- Each expert gives his answers to the questions in an independent and anonymous way.
- The responses of each expert are analysed by the monitoring team. The lower 25% and the upper 25 % of responses are excluded.
- The set of responses is then sent back to experts and they are asked if they wish to revise the initial predictions.
- The process is reiterated until a degree of consensus is reached by experts.

The Delphi method has undergone many variations. One of the most important was letting the experts indicate their own expertise for each question (for example rating their expertise on a scale of 1 to 7). This variation was supposed to improve accuracy, since only opinions of experts with "higher" expertise were used to determine the distribution of opinions for that item. This approach was challenged when it was found that self-rating of participants did not coincide with "objective expertise". Moreover, it has been found that women consistently rate themselves lower than men.

The Delphi method has been criticised as well as supported. Major criticisms to the method were:

- A low level reliability of judgements among experts and therefore dependency of forecasts on the particular judges selected;

- Sensitivity of the results to ambiguity in the questionnaire that is used for data collection in each round.
- Difficulty in assessing the degree of expertise incorporated in the forecast.
- Responses can be altered by monitors in the attempt of moving the following round of responses in the desired direction.

Nevertheless, it must be acknowledged that there has been many poorly conducted Delphi applications and it is not correct to equate the applications of the Delphi method with the Delphi methods itself. There is in fact an important conceptual distinction between evaluating a technique and evaluating an application of a technique.

It can be concluded that in general the Delphi method is useful in answering one specific and single-dimension question. There is less support for its use to determine complex forecasts concerning multiple factors since the collation of expert judgements suffers from the possibility that interactions between forecasted items may not be fully considered. Delphi applications seem to have dramatically decreased and not to influence appreciably contemporary discussions of expert opinion.

An improvement in forecasting reliability in the Delphi method consisted in taking into consideration the possibility that the occurrence of one event may change the probability of occurrence of other events included in the surveys. Cross impact analysis was therefore developed as an extension of Delphi method.

2.3 Cross Impact Analysis

As described above, the basic limitation of the Delphi method is that it only provides separated forecasts, that is to say events and trends are considered separately from each other.

Cross impact concept originated with Gordon and Helmer in 1968 in conjunction with a forecasting game called *Future* developed for Kaiser Aluminium and Chemical Company (Gordon, 1994). It represented an effort to extend the forecasting technique of the Delphi method. In 1968 at UCLA, Gordon and Hayward programmed the approach and developed a computer-based tool to cross-impact matrix analysis. In this approach, events were recorded on an orthogonal matrix and at each matrix intersection the question was asked: "*if the event in the row were to occur, how would it affect the probability of occurrence of the event in the column*?". The judgements were entered in the matrix cells. Cross impact analysis attempts to reveal the conditional probability of an event given that various events have or have not occurred.

The major steps in the use of cross-impact analysis for evaluating future situations can be summarized as follows:

- Define the events and trends to be included in the analysis.
- Define the planning interval and subintervals, "scenes".
- Develop cross-impact matrices to define the interdependencies between events and trends.
- Estimate the entries in the cross-impact matrix, i.e., information on how the occurrence of an event E_i or how the deviation of a trend T_j from its expected value in a given scene would affect other event probabilities and trend values in later scenes.
- Estimate the initial occurrence probabilities of each event in each scene.
- Estimate the value of each trend at the beginning of each scene.
- Perform a calibration run.
- Define tests and actions to be run with the matrix.
- Perform the cross-impact calculations.
- Evaluate results

The initial occurrence probabilities of events, values or trends, and the magnitude of impacts between the variables may be estimated by individual experts but more commonly estimated by groups containing experts from the various disciplines covered by the events. Delphi questionnaires or interviews also can be used to collect these judgements.

Since its first applications, expert judgement has been used in a more or less systematic way in many fields (see chapter 7). Different approaches have been used but uniformity in its use is still lacking.

3 FORMAL EXPERT JUDGEMENT

Several formal approaches to expert judgement exist for application in different fields (Bolado and Gallego, 2000). Some examples are the approach developed by the Stanford Research Institute in the 70s (Merkhofer, 1987), the approach used in “Severe Accident Risks: An Assessment for Five U.S. Nuclear Power Plants”, NUREG-1150, USNRC 1990 and the approach developed by the University of Delft (Cooke and Goossens, 2000).

It is not in the purpose of this document to analyse the different approaches that have been developed and used. The document aims at providing a general structure for formal expert judgement, based on the above-mentioned methods.

The goal of applying a formal or structured expert judgement approach is to enhance a rational consensus. Principles representing an attempt to formulate guidelines for the use of expert opinion have been proposed by Cooke (1991) as the following:

- *Reproducibility*: all data and all processing tools are open to peer review and the results must be reproducible by competent reviewers.
- *Accountability*: the source of expert subjective probabilities must be identified.
- *Empirical control*: expert probability assessments must in principle be susceptible to empirical control.
- *Neutrality*: the method for combining/evaluating expert opinions should encourage experts to state their true opinion.
- *Fairness*: the experts are not pre-judged and they are treated equally prior to processing the results of observations.

Experts may be asked for judgements in different forms (Cooke and Goossens 2000). In early methods, e.g. the Delphi method, experts were asked to guess values of unknown quantities, and the answers were thus *single point estimates*. One disadvantage of such estimate is that it does not give indication of uncertainty. One form of expert judgement is to ask experts to rank alternatives with *paired comparisons*. A third form of judgements is *discrete event probability*, where experts are asked to assess the probability of an uncertain event in the form of a single point value in the [0,1] interval. The most important form of expert judgement is the *distribution of continuous uncertain quantity*. We focus in this report to this type of expert judgement.

The process of formal expert judgement usually consists of the following steps:

1) Identification and selection of issues about which the expert judgements should be made.
2) Identification and selection of experts.

3) Training of experts and definition of variables to be elicited.
4) Individual work of experts.
5) Elicitation.
6) Analysis and aggregation of results and, in case of disagreement, attempt to resolve differences.
7) Documentation of results, including expert reasoning in support of their judgement.

Various references (see e.g. Otway and Winterfeldt, 1992, Cooke and Goossens, 2000) have a slightly different list of phases, but essentially the procedure is similar.

The first step, identification and selection of the case to be studied, includes collection and preparation of background material. This documentation, which corresponds to the *case structure document* and the *elicitation format document* defined by Cooke and Goossens (2000), should contain the frame for the expert judgements, specifying issues that should be taken into account, and a preliminary definition of the variables to be elicited. This documentation will be reviewed and refined later by experts. When the variables to be elicited, i.e. *query variables*, are defined, the following rules should be applied (Cooke and Goossens 2000):

- Ask for values of observable or potentially observable quantities.
- Formulate questions in a manner consistent with the way in which an expert represents the relevant information in his/her knowledge base.

Note that the first rule implies that one should try to avoid, if possible, asking experts to judge uncertainties of a probability, since experts could find extremely difficult to provide uncertainty ranges for a probability. Thus it is advisable to formulate the questions in terms of experiments whose measurable results are physical magnitudes.

The second step is the identification and selection of experts. Selection criteria for experts have been set e.g. in the "*Procedures Guide for Structured Expert Judgement*" (Cooke and Goossens 2000). These criteria are:

- Reputation in the field of interest.
- Experimental experience in the field of interest.
- Number and quality of publications in the field of interest.
- Familiarity with uncertainty concepts.
- Diversity in background.
- Awards.
- Balance of views.
- Interest in the project.
- Availability for the project.

An important issue is to try to guarantee as much as possible independence among experts. A way to achieve this objective, at least partially, since it is almost impossible to achieve it completely, is to look for as much diversity as possible in the set of experts, seeking diversity in education and in professional background among others.

Besides the technical or *substantive experts* (also known as *domain experts*), the expert judgement process should involve so called *normative experts* that should be knowledgeable in the field of subjective probability. Their task is to conduct the expert judgement process, and they are responsible for the training, elicitation, aggregation and final reporting of the case. For clarity, **substantive expertise** refers to the knowledge that the assessor has about the quantity to be assessed; **normative expertise** refers to skills of the assessor in expressing his or her beliefs in probabilistic form. Additionally we could also consider '*generalists*'. Such experts have wide background knowledge in the area of interest, have a very clear idea of the target of the whole project, and are competent on that area of knowledge. They can provide a lot of help to both normative experts and substantive experts, e.g. providing information and helping the interaction between normative and substantive experts.

Another issue is the ideal number of substantive experts to collaborate in the project. Some Bayesian argumentations arrive at the ideal range of 3 to 5 experts (Clemen and Winkler, 1985). This is based on the fact that, under even quite non-realistic conditions of independence between experts, the reduction in uncertainty obtainable reaches an asymptotic behaviour (no improvement of accuracy) for such small numbers. Lack of experts in some areas and budget restrictions do also force such numbers.

The expert judgement process can also be guided, depending on available resources, by a facilitating team composed of normative experts, substantive experts and recorders (USNRC, 2004).

Once the case study and elicitation format have been defined, a *dry run exercise* can be performed. The dry run exercise aims at analysing the case structure document and the elicitation document to find out all ambiguities and misunderstandings related to the case and query variables in order to remove them. One or two persons experienced in the field of interest should be asked to provide comments on the two documents. The dry run experts should preferably come from outside the selected panel of experts. In case this is difficult to be realized, expert panel members may be asked to do the dry run. After the dry run exercise the case structure document and the elicitation format document will be finalised and sent to the expert panel.

In order to familiarise experts with the process of providing subjective assessments and understanding subjective probability related issues, a training

_____ should be given. There should be a clear definition of the issues on which experts have to make judgements and, as a help, decomposition can be used in case of complex issues. The training of experts is discussed in more detail in Chapter5.

After the meeting, experts should perform their own independent analyses. Depending on the subject to be assessed, the analyses may need time between a few hours and a week per expert (Cooke and Goossens 2004). In their independent analysis, experts can use all their available knowledge and sources of information. However, they will have to report on it when documenting their analysis.

The experts are invited to an elicitation session. In the European Guide for Expert Judgement (Cooke and Goossens 2000), it is recommended that experts be elicited individually. In these sessions, all results are reviewed and discussed. In approaches used by the NRC (NUREG-1150 1990 and Bonano et al. 1990) the experts discuss their analyses in a joint meeting, but without expressing their numerical judgements. Thus the experts have a possibility to compare their problem decomposition, events and variables, and ask further questions related to the issue. After this session, the experts are interviewed individually to obtain their estimates. In the elicitation, the normative expert asks the specialists about their reasoning, ensures that the required information is obtained, checks the consistency of judgements especially with the laws of probability and documents the numerical results for later processing. The most common approach to elicit the distribution of the uncertain quantity is to ask the expert a judgement on the median value and some fractiles of the distribution. It is also possible to ask parameters of some distributions.

The numerical results obtained from experts in elicitation sessions are finally analysed and aggregated. There are several possible methods for aggregation, and some of them are summarised in Chapter 6. In connection to the aggregation, also sensitivity analyses should be performed in order to study the sensitivity of the final distribution to the judgements of various experts. Calibration and weighting of experts are also discussed in Chapter 6.

The last step in the formal expert judgement process is the documentation of the study. This documentation should include the description of the case, individual reports of experts describing their reasoning and analyses, and description of the aggregation of the judgements and the final result of the study. The documentation should meet the objectives of the formal judgement process, i.e.:

1) To improve decision-making.
2) To enhance communication.
3) To facilitate peer review appraisal.
4) To recognise and avoid biases in expert judgement.

5) To indicate unambiguously the current state of knowledge about important technical and scientific matters.

6) To provide a basis for updating knowledge.

4 BIASES

According to the subjectivistic school of probability, the probability of an event is a measure of a person's degree of belief that the event will occur. Probability is not a property inherent to the event, but a statement of an observer's judgement that it will occur. At the early stage of the use of subjective probability it was believed that these constructs existed in the heads of subjects. However it has become apparent that in most cases experts must synthesize or construct probability values and distributions when an analyst asks for them. In this process of estimating probabilities or determine degree of belief, experts use "rules of thumb", the so-called ***heuristics***.

Heuristics are easy and intuitive ways to deal with uncertainties, but since they are at best only approximate procedures, they can lead to predictable "errors". By "error" we mean a violation of the axioms of probability or an estimate that is not in accord with the expert's beliefs and that the expert would like to correct if the matter was brought to his/her attention. 'Errors' could also be systematic underestimation or over estimation of quantities. These "errors" in the context of expert judgement are called ***biases***. Because of the existing biases of the eliciting heuristics the question of how to minimize biases and systematic errors in elicitation is essential.

Awareness of heuristics and biases may help individuals to make better probability assessments, and thus they should be introduced and discussed during the training for expert judgements.

4.1 Cognitive Biases

Biases are linked to the way expert processes the information that is to say to the way the expert reasons. This depends on the rational and experiential level of the expert, called cognitive-experiential self-theory. ***These are cognitive biases*** and they are a distortion of the way we perceive reality.

Biases linked to following heuristics are discussed in this chapter:

- Representativeness
- Overconfidence
- Availability
- Adjustment and anchoring
- Lack of capability to deal with some statistical concepts
- Control

For each of these heuristics, several biases are presented. We shortly summarise the main biases. For a more detailed description and examples, see Kahneman et al. (1982).

Representativeness

Granger Morgan and Henrion (1990) suggests that in judging the likelihood that a specific object belongs to a particular class of objects, or that an event is generated by a particular process, people expect the structure or details of the object or event to reflect the larger class or process. This can be shortly summarised by saying that people tend to think that if X represents well group A, the probability that X belongs to A is high.

People expect that a sequence of events generated by a random process will represent the essential characteristics of that process even when the sequence is short. For example, people judge the string of coin tossing HTHTTH to be more likely than either the string HHHTTT or the string HTHTHT because they know that the process of coin tossing is random. The three sequences are equally likely to happen, however the first string appears more random that the other two outcomes. This is due to the fact that people tend to expect in the small behaviour, the same behaviour that one knows exist in the large.

This is what Kahneman et al. (1982) have termed "belief in the law of small numbers" and is frequently evidenced even among technical people with substantial formal statistical training. This bias is called *misconceptions of chance.*

Another consequence of representativeness is that people often pay too much attention to specific details while not paying enough attention to base rates. Often the relative frequencies of the groups are neglected. A clarifying example is provided by Kahneman et al. (1982):

> "A panel of psychologist have interviewed and administered personality test to 30 engineers and 70 lawyers, all successful in their own fields. On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written. …For each description, subjects were asked to indicate the probability that the person is an engineer, on a scale form 0 to 100. It was found that responses were based on how much the described person was judged to sound like an engineer or lawyer without regard for the 30:70 ratio. As an example, the description: Jack is 45-year-old man. He is married and has four children. He is generally conservative, careful and ambitious. He shows no interest in political and social issues and spends most of his free time on his hobbies which include home carpentry, sailing and mathematical puzzles, was judged with very high probability to involve an engineer, while the description "Dick is a 30-year old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues" was judged at 50:50, not 30:70, because subjects clearly recognized that the description said nothing relevant about lawyer-engineer distinction and appeared to forget the base rate information."

When no specific evidence is provided, the prior probabilities are properly utilized, when worthless specific evidences is given, prior probabilities tend to be ignored. This bias is called *insensitivity to prior probability of outcomes*.

The representativeness heuristic also leads people to ignore effects due to sample size. The size of a sample withdrawn from a population should affect the likelihood of obtaining certain results in it. However, sample sizes are ignored, and people tend to use only superficial similarity measures. A clarifying example presented in Kahneman, et al. (1982) is the following:

> "A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50 % of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50 %, sometimes lower. For a period of 1 year, each hospital recorded the days on which more than 60 % of the babies born were boys. Which hospital do you think recorded more days?"

The results showed that of the 95 interviewed subject, 21 opted for the large hospital, 21 for the smaller hospital and 53 thought both hospitals recorded about the same number. In contrast, sampling theory implies that the smaller hospital is much more likely to see more than 60% boys on any given day. As a rule of thumb, larger absolute deviations from mean values are expected in larger populations, while larger fractional deviations from mean values are expected in smaller populations. This bias is called *insensitivity to sample size.*

Note: Regarding the births as independent tosses of coin with probability *p* of "heads" (for boys), the variance of the proportion S_n/n of heads in n tosses is p(1-p)/n. Since this variance decreases with n, the probability of a given relative deviation from the mean value (p) decreases with n.

Availability

In *availability heuristic* people estimate the probability of an outcome based on how easy that outcome is to imagine. That is to say, their probability judgement is driven by the easiness with which they can think of previous occurrence of the event, or the ease with which they can imagine the event occurring. Also, vividly described, emotionally-charged possibilities will be perceived as being more likely than those that are harder to picture or are difficult to understand, resulting in a corresponding bias.

An example from Granger Morgan and Henrion (1990) reports that in estimating the likelihood of encountering a highway patrolman on the way to work, one thinks about how often he/she has encountered patrolmen during the daily drive to work in the last ten years. The result would be in general quite good estimate. The ease with which one is able to think of previous encounters with patrolmen is likely to be well correlated with the actual frequency. However, some events or items may be much easier to recall than others, therefore biases in the estimation can be easily introduced.

When the size of a class is judged by the availability of its instances, a class whose instances are easily retrieved will appear more numerous than a class of equal frequency whose instances are less retrievable. An elementary

demonstration (Kahneman et al. 1982) is the case in which subjects heard a list of well-known personalities of both sexes and were then asked to judge whether the list contained more names of men or women. Different lists were presented to different groups of subjects. In some lists the men were relatively more famous than women, in other the opposite. In each of the lists, the subjects erroneously judged that the sex (class) that had the more famous personalities was the more numerous. This bias is called *retrievability of instances*.

Sometimes there is the need to assess the frequency of a class whose instances are not stored in memory but can be generated according to a given rule. In such situation, one tends to generate instances and evaluate their probabilities by the ease with which the relevant instances can be constructed. Consider the following example for clarification (Kahneman et al. 1982):

> "Consider a group of 10 people that form committees of k members, with k value between 2 and 10. How many different committees of k members can be formed? When the question was posed to subjects, the median estimate for 2 was 70 and the median estimate for 8 members was 20. "

In reality there are as many committees with two members as committees with 8 members (correct answers is 45). The responses indicate a bias due to the fact that it is much easier to imagine committees of two members and to imagine grouping ten people into different groups of two, than it is to imagine groups of eight. Committee of two are therefore more available. This bias is called *ease of imaginability*.

Another example of the operation of the heuristic of availability involves the perception of risk. When asked to estimate the probabilities of death form various causes, subjects typically overestimates the risk of well publicized though less frequent causes (botulism, snake bite) and typically underestimate less publicized causes (stomach cancer, heart disease). In this case the affect heuristic (discussed later in the text) plays also a role since remembered images come marked with affect. An explanatory example is presented in figure 1.

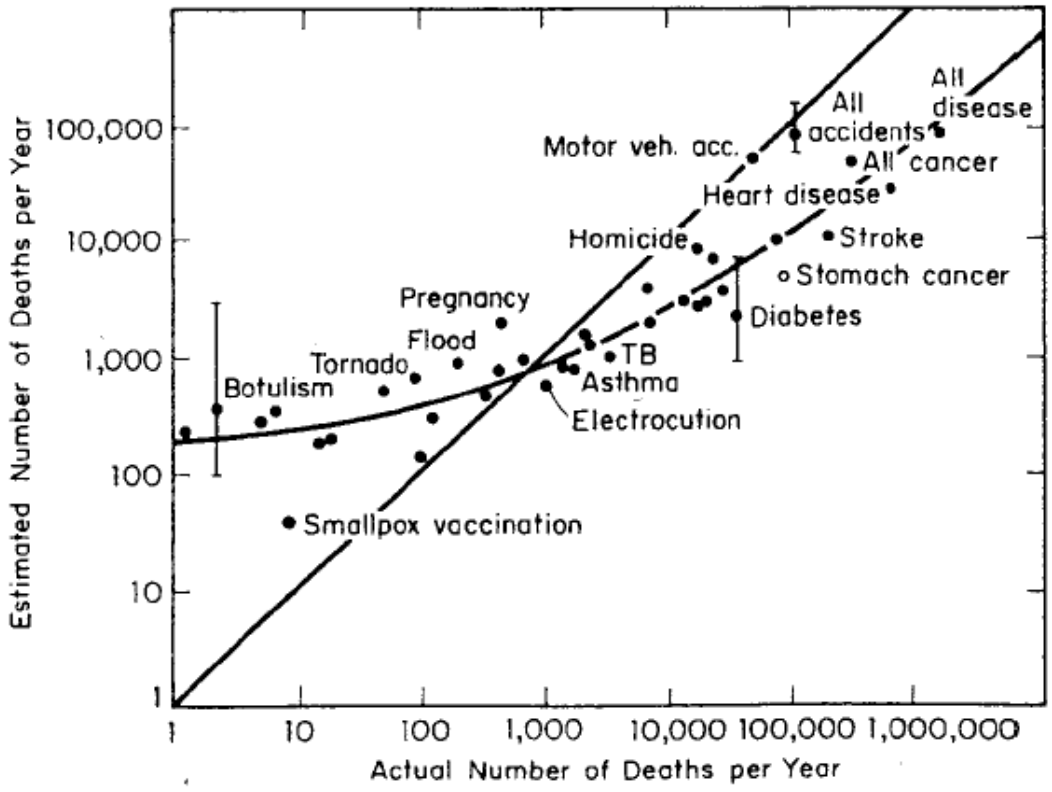


Figure 1: Plot showing the geometric mean of people's estimates of the annual numbers of deaths form 41 causes (vertical axis) versus the actual numbers of deaths (horizontal axis). If judged and actual frequency were equal, the data would fall on the straight line. In general, the occurrence of frequent causes of deaths is underestimated and that of less frequent causes is overestimated. The operation of bias form the heuristic of availability is clearly illustrated by the points for stroke and botulism. As an index of variability across individuals, vertical bars are drawn to depict 25th and 75th percentiles of the judgements for botulism, diabetes and all accidents. The figure is drawn from Slovic, Fischoff and Lichtenstein, 1982. (In Granger Morgan and Henrion, 1990)

Biases through the use of availability can also arise because of variations in the ease with which an event can be imagined. This especially applies in the context of scenarios in which people often assess the probability of occurrence of a situation that links several events in sequence. In this situation the judgement of how frequently events co-occur is linked to the strength of the associative bond between them. Linked events in a scenario are easier to imagine that the individual events in isolation. When the association is strong, judges are likely to conclude that the events have been frequently paired. Therefore, strong associates will be judged to have occurred together frequently. This bias is called *illusory of correlation.*

In conclusion, use of availability heuristic will yield reasonable results when a person's experience and memory of observed events corresponds fairly well with

actual event frequencies; is likely to lead to overestimates if recall or imagination is enhanced (e.g. recent experience, dramatic or salient events, plausible scenario, etc.); and is likely to lead to underestimate if recall or imagination is difficult (e.g. no recent experience, concept abstract, not encoded in memory, etc.).

Anchoring and adjustment

Anchoring and adjustment is a psychological heuristic said to influence the way people estimate probabilities intuitively. Under this heuristic a natural starting point, or anchor, is selected as representing a first approximation of the value of the quantity to be estimated. This value is then adjusted to reflect supplementary information. Typically the adjustment is insufficient and the result is biased toward the anchor. Kahneman et al. 1982, report a number of experimental examples of this bias.

> "In one experiment subjects were told the objective was to estimate a quantity, Q, in percent (for example the percentage of African countries in UN). A "wheel of fortune" was spun for the subject to produce a quantity A, with 0<A<100. Subjects were led to believe that the result was a random number between 0 and 100, although the wheel always yielded either 10 or 65. After obtaining A, subjects were asked if Q was greater or less than A. Subjects were then asked to estimate Q by adjusting their response up or down form A. When A was 10, the median estimate of Q was 25, but when A was 65, the median estimate of Q was 45. "

This bias is called *insufficient adjustment.*

Biases in the *evaluation of conjunctive and disjunctive events* are explained as effect of anchoring. Studies on judgement of probability indicate that people tend to overestimate the probability of conjunctive events and to underestimate the probability of disjunctive events. The stated probability of the elementary event (success at any one stage) provides a natural starting point for the estimation of the probabilities of both conjunctive and disjunctive events. Since adjustment from the starting point is typically insufficient, the final estimates remain too close to the probabilities of the elementary events in both cases. The overall probability of a conjunctive event is lower that the probability of each elementary event, whereas the overall probability of a disjunctive event is higher than the probability of each elementary event. As a consequence of anchoring, the overall probability will be overestimated in conjunctive problems and underestimated in disjunctive events.

Overconfidence and calibration

Overconfidence is one of the most important biases. Overconfidence is not directly associated with an estimation heuristic, however in some cases anchoring can aggravate it. When assessing probabilities, overconfidence expresses itself as poor calibration.

Calibration can be measured empirically in experiments that involve many assessments of quantities about which assessors have some relevant but imperfect knowledge, and whose true value can be found by the experimenter (*discrete tests*). The assessed quantities are plotted against the true value in what is called *calibration curve*. For a well-calibrated judge the curve should be near the diagonal (figure 2). For an under confident judge the assessed probabilities are nearer 0.5 than they should be; more typically judge are overconfident and the probabilities are assessed too near certainty (0 or 1).

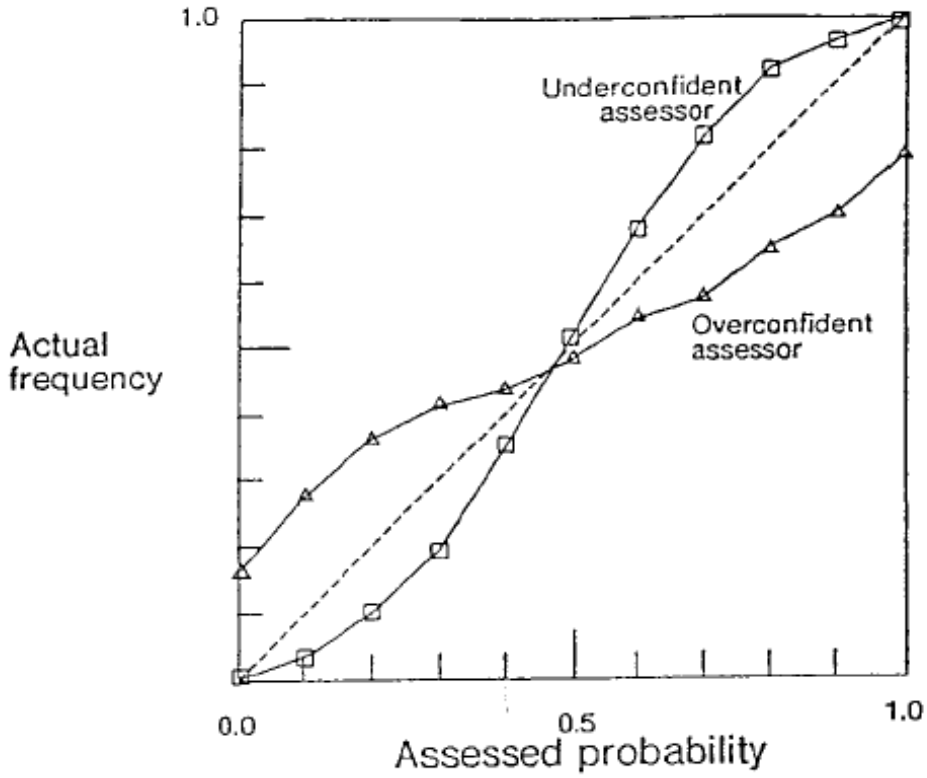


Figure 2: Calibration curves from assessments of discrete probabilities (Granger Morgan and Henrion 1990)

Calibration curve can also be compiled for assessments of continuous distributions for unknown quantities (*fractile or quantile tests*). In this case the fractile of the assessed distribution at which the true value occurs is recorded. These fractiles form a distribution from 0 to 1 and their cumulative distribution is a calibration curve (figure 3).

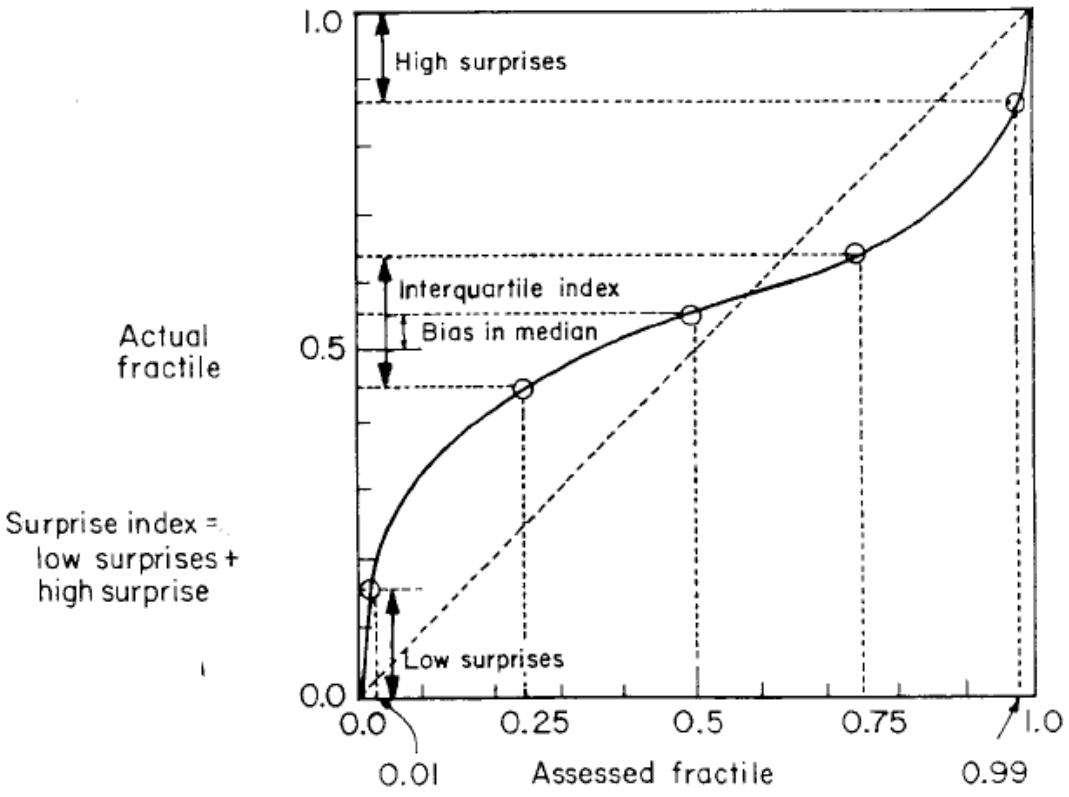


Figure 3: Calibration curve from assessments of continuous distributions (Granger Morgan and Henrion 1990)

The curve of a well-calibrated subject would be the diagonal. Supposing that we ask an expert for his 1%, 25%, 50, 75% and 99% quantiles for a large number of variables for which it is possible to know the value later. In case of well-calibrated expert, it should be expected that about 1% of the true values fall below the 1% quantiles, roughly 24 % should fall in 1%-25% as well as 75-99% intervals. The *interquartile range* is the interval between the 25% and the 75 % quantiles. 50% of the true values should fall within the interquartile range.

The *surprise index* is the percentage of true values that fall below the lowest or above the highest quantiles. Usually it is defined as the proportion of values falling outside the 98% credible interval, that is, less than 1 % fractile or greater than the 99% fractile.

A perfectly calibrated expert would have an interquartile index of about 50 % and a surprise index of about 2%.

Lack of capability to deal with some statistical concepts

A quite frequent problem, even among people with some statistical background, is the difficulty to distinguish means and medians of random variables. Experiments show that, in many cases, when subjects are asked for means, they usually provide medians. This is probably due to the additional complexity of estimating means (computing an integral in the continuous case and a sum usually of many terms in the discrete case). Spread statistical measures are also difficult to be evaluated, i.e. variances, which is also linked to the problem of overconfidence. Bayes' theorem is another probability tool not frequently understood and appropriately used by subjects to update information.

Winkler (1967) has also reported about the tendency of subjects with some statistical background to fit their opinions to the normal model, which seems to be related to the pervasive use of this probability model in basic probability and statistics courses. Hogarth (1975), nevertheless, defends the idea that this is an uncertainty reducing mechanism of people that prefer to think in even situations instead of odd ones (symmetry is one of the properties of the normal model).

Control

Subjects tend to act in certain case as if they can influence the situation over which in reality they have no control at all. This may cause distorted results in probability assessments. An example can clarify this case: in a game consisting of two opponents having to cut a deck of cards, participants were asked to bet with one of the two opponents. One opponent was instructed to be shy and insecure, and the other opponent was instructed to be confident and self-possessed. It was conjectured that the subjects would think they had a better chance of winning against the insecure opponent since the amount of money they were willing to bet against him was higher.

4.2 Other biases

Biases can also be introduced by the methods used to structure the expert judgement process. ***These are structural biases***. Structural biases can also occur when experts are unduly influenced by the way a problem has been structured before it is presented to them (Otway and Winterfeldt 1992). Probability judgements can be influenced by the organisation of events, since analysts are relatively insensitive to omitted events and overly influenced by events that have been presented in great detail.

Further, there are biases linked to the distortions of judgement due to expert's ideology and believes. In this situation judgements are distorted willfully (as in lying). ***These are motivational biases.*** Motivational biases can occur also because the expert has a stake in the issue considered that may lead to

conscious or unconscious distortions of his judgements (Bonano et al. 1990). For such biases, awareness of motivational factors is important.

Finally, biases can be also due to the "experience" of the subject that therefore tends to judge based on his background. ***These are background biases*** (i.e. what an individual might see as reasonable or would expect based on his background). These biases are natural, but it is important to get the subject to consider them.

In a recent article, Slovic et al (2004), reports on diverse set of empirical studies in support of "**affect heuristic**". Affect in this context means the specific quality of "goodness" or "badness", experienced as a feeling state (with or without consciousness) and demarcating a positive or negative quality of a stimulus. Biases in probability and frequency judgements that have been attributed to the availability heuristic may be due, at least in part, to affect. Availability may work not only through ease of recall, but also because remembered images come marked with affect (consider for example the cause of death in which there is an overestimation for highly publicized one and an underestimation for not publicized one shown in figure 1).

4.3 General remark on biases

Most of the experimental results available in scientific literature about biases are based on studies done with students providing answers to general culture questions. This is why these results are considered with caution by some authors. In fact, some relevant authors like Lindley (1988) are not surprised about doubts some psychologists have about people as probability assessors, mainly if their capability is estimated through questions like 'What's the probability that the number of telephones in Ghana is larger than 100,000?'

Mullin (1986) did a series of studies with experts in the areas of electromagnetic fields and hydrology, getting opinions about their respective areas of expertise and about general culture. The results of these studies showed a large difference between the assessments they did when they worked as experts and when they worked as non-experts. When they worked as experts they were quite more careful providing estimates, gathering information, identifying uncertainty sources and building models. Two of the main conclusions of these studies were that the results of experiments done with non-experts could not be directly extrapolated to experts, and experts were usually less overconfident than normal subjects when providing their opinions. Awareness of biases by formal training is probably the best that can be done to avoid them, instead of applying some not always well-justified debiasing techniques.

5 TRAINING

Ideally the experts should have a solid background in probability theory and statistics. However, this is often hard to achieve, especially if probabilistic and statistics are not used daily in their work. Even if the experts are familiar with most of the concepts, they may lack the knowledge of subjective interpretation of probability and may not be aware e.g. of the cognitive biases related to judgements. Thus a training session is an important part of the expert judgement process.

The expert training should cover the following issues (see e.g. Bonano et al. 1990):

- Familiarising the experts with the expert judgement process and motivating them to provide formal judgements.
- Giving the experts practice in expressing their judgements formally.
- Informing the experts about possible biases in expert judgement and the application of debiasing techniques.

In connection to the training session, the capability of experts in making probability judgements can be evaluated through exercises such as the development of set of questions. If considered necessary, a calibration of experts can be made based on their answers to these test questions.

5.1 Introduction to expert judgement and basic concepts of probabilities

Providing formal expert judgements is usually unfamiliar to experts. Further they may worry that their judgements may be misused or misinterpreted. Thus it is very important to familiarise experts with the process. The need and purpose of expert judgements should be made clear, and it should be stressed that there is not only one right answer. The formal expert judgement is rather a tool to summarise the current information, and it identifies where sufficient knowledge exists and where more research is needed.

Since expressing judgements as probabilities is seldom part of daily life of experts, it is useful to explain basic concepts and main properties of probabilities during the training. Use of expert opinions to produce probability distributions to express the uncertainties is based on the concept of subjective probability. Thus it is very important to explain the various concepts or interpretations (e.g. classical, frequentistic and subjective) of probability.

5.2 Decomposition

Experts should be trained to some extent to explicitly express their judgements. Making implicit judgements explicit can be helped with practical examples. Most

expert judgements can be aided by decomposing the problem (disaggregation), and examples of decomposition can be helpful. Problem decomposition is widely used in scientific studies to simplify a complex problem into components that are more manageable and more easily solved. These less complex assessments are then recombined into a probability distribution for the quantity of interest. Bonano et al. (1990) provide a good description of decomposition techniques and discuss advantages and disadvantages of decomposition.

Examples of modes of decomposition are event trees, fault trees and functional decompositions. Fault trees focus on a possible failure of the system and traces back the possible causes of this failure at component level. In the event tree technique, the analysis is started from an initiating event, and the probabilities of successive events are conditional on their predecessors. Decomposition may also use physical models of the phenomena. In such case the physical relationship between the quantity of interest and several constituents is expressed through a mathematical function.

Decomposition can also be used less formally. The goal may be to promote deeper insight into the rationale for judgements and to enhance the interchange of beliefs and assumptions about the likely causes of studied events without formally encoding the decomposition (Bonano, Hora et al. 1990).

Individual experts may be allowed to choose their own decomposition, or it may be decided that all expert should use the same consensus decomposition. Both approaches have their advantages and drawbacks. The advantage of multiple decompositions is that a wider variety of approaches to solve the problem is permitted.

5.3 Training on biases and debiasing techniques

It is very important to get experts to think carefully about substantive details of each judgement they make. Studies presented in Granger Morgan et Henrion (1990) report improvement in calibration when experts were asked to provide lists of reasons justifying their judgements as opposed to just providing the judgements. In other studies experts were asked to list either one reason for their response or one reason against their response, or list one reason for and one against. Results of the studies show that only in the case of one reason given against their choice (*disconfirming information*) was overconfidence significantly reduced.
However, further studies have shown that debiasing by asking for reasons may have more impact on judgement tasks for which the expert has limited experience than for tasks with which he or she is intimately familiar.

Training should also be provided on the heuristics and on the biases they lead to. As already underlined in Chapter 4, training on biases may help individuals to

make better probability assessments. Their knowledge is therefore essential in the elicitation process to avoid systematic errors. Detailed description on biases and their mechanisms was given in Chapter 4.

5.4 Exercises

It is common in the expert judgement process to have exercises for expressing uncertainty with probabilities. These exercises do not have necessarily to be related to the area of expertise of the experts.

Cooke and Goossens (2000) support the use of *calibration* or *seed variables* to be posed to expert prior to the elicitation process. Seed variables are variables whose values are or will be known within the framework of the exercise to the normative expert, but not to expert. Seed variables are important for assessing the performance of the combined experts' assessment. They are also an important feedback to expert that can help them to estimate their subjective sense of uncertainty.

Typically, experts are asked to provide 5%, 50% and 95% quantiles for the distributions of so-called seed variables. Let us assume that we have a set of seed variables, and experts provide their uncertainty distributions on these variables. Afterwards, the true values are shown and the performance of the experts can be evaluated. In principle, events that are assigned a given probability should occur with a relative frequency equal to that probability. For example, if we have a set of 20 seed variables, for a well-calibrated expert, approximately one out of 20 true values should fall below the estimated 5% quantiles, and one over the 95% quantiles. In 10 cases the true values should be larger than the expert's median and in 10 cases smaller. Comparing the true and the estimated values, the experts can identify whether they tend to be e.g. overconfident, or give systematically too high/low values. If 3 or 4 values fall outside the 90 % bands, it can be interpreted as sampling fluctuations, but if e.g. 10 out of 20 true values are outside the bands, there is reason to suspect that the expert chooses the uncertainty bands too narrowly.

However, it should be mentioned that it is not clear whether the training on trial tasks will improve the actual performance of the experts in their assessment work since different studies show contrasting results. For example moderate increase in the *interquartile index* and reduction in *surprise index* (see figure 3 in chapter 4) have been reported in cases of continuous distribution assessment, on the other hand considerable improvement in calibration was found in discrete choice tasks with different level of difficulty (Granger Morgan and Henrion 1990).

The information obtained from exercises with seed variables could also be used to evaluate the expert performance for weighting the experts. The performance measures and weighting of experts is discussed in section 6.2.

6 ELICITATION AND AGGREGATION OF JUDGEMENTS

6.1 Elicitation techniques

Elicitation, or probability encoding, is the process of extracting and quantifying the subjective judgements about uncertain quantities. Accepting the fact that often people do not have clear intuitions about probabilities and find it easier to express probabilities quantitatively, some aids can be given in the elicitation process.

As an aid to conceptualising probabilities, some researchers have suggested assessors to visualize an urn containing coloured balls in proportions that approximate the required probabilities. However it is not clear if this can help people to relate probability to their everyday experience. Other aids in the elicitation process can be the use of methods that favour indirect responses mode in which expert can avoid implicit mention of probabilities, such as the "reference lottery" and the "probability wheel". More details on these two methods are in Granger Morgan and Henrion (1990).

It can be concluded that in general physical aids can help in visualizing probability and these can be very useful even for subjects with rather considerable experience.

Under some circumstances, experts do not feel comfortable with the usual probability scale. In those cases alternative scales could be sought, as for example *odds* and *log-odds.* An odd is the quotient between the probability of an event and its complementary, so if *P* is the probability of an event, its odd is

$$odd = \frac{P}{1-P}$$

A *log-odd* is the logarithm of an *odd. Odds* are suitable for subjects that prefer to express relative probabilities between complementary events, as for example, 'for this parameter a value smaller than *a* is 100 times more likely than a value larger than *a*'. The corresponding *log-odd* is 2.

Another aid to assess probabilities is the jargon of gamblers. Some subjects find easy to express probabilities as bets like "I bet *h* against *k* in favour of this option", which implicitly means that the subject assigns probability *P=h/(h+k)* to the event under study.

Bonano et al. (1990) provide a good summary of main classes of procedures and the nature of questions asked in the elicitation. The elicitation techniques are classified according to whether it is question of magnitude judgements about events or indifference judgements about gambles. Further, classification is made according to variables, whether they are discrete events or continuous quantities.

It is stated in Bonano et al. (1990) that it is important to begin with easy questions, and that it is preferable to select observable quantities for eliciting probabilities. It is also useful to ask the same question in different ways and use the results for consistency checking.

The *fractile technique* is the most commonly used technique to elicit continuous uncertain quantities. It is used to construct the cumulative distribution function of the uncertain quantity. The experts may first be asked the lower and upper bounds, i.e. 0% and 100% fractiles. Assessment may also focus on 1% and 99% or/and 5% and 95% fractiles instead of the absolute maximum and minimum values. After having obtained the extremes, the normative experts ask for the median of the distribution. Further, 25% and 75% fractiles are commonly assessed.

In the *interval technique*, the normative expert preselects points of the uncertain quantity and asks the specialist to assign probabilities to intervals defined by them. First, extremes are asked in the similar way as in the fractile technique. Then normative expert chooses three to seven points, possibly equally spaced, from this interval. In the *open interval technique*, experts assign for each point probabilities that the actual magnitude falls below or above of this point. In the *closed interval technique*, experts assign probabilities that the true magnitude falls in each interval defined by the points. Any of the aforementioned techniques to aid probability statements (reference lottery, reference urn, probability wheel, odds, lo-odds and bets) may be of help to some experts when using the interval technique.

Some experts with skills in probability theory and statistics could feel confident enough to provide their full characterisation of uncertainties as probability density functions or probability distribution functions (saying, for example type of distribution and parameters), though this is not a frequent case. In this case, the analysts should design questions to check if the expert fully understands the meaning of the distributions provided. Questions to check consistency are also advisable.

Hampton et al. (1973) report about Smith's method, called psychometric classification. The expert must provide minimum and maximum values for the variable under study. The analysts divide this range in a set of non-overlapping segments that cover the whole range and ask the expert to sort those segments from the most to the least likely. Additionally the expert is asked to sort from the largest to the smallest the differences between contiguous segments. These data in addition to the use of a procedure suggested by Kendall allows creating a histogram for the parameter under study. Experimental studies suggest that this is a reliable method that provides distributions with larger variances than other techniques. The surprising fact of this technique is that experts are never asked for probabilities.

Some Bayesian techniques like the equivalent prior sample (EPS) and the hypothetical future sample (HFS) developed by Winkler (1967) have not been used widely and were found difficult to apply by experts and by its very developer.

An important principle is that the experts should be allowed to choose the scales, techniques and aid tools they prefer. The more comfortable they feel when providing their estimates, the better.

6.2 Expert weighting

Experts may not be equally good in providing judgements. This could in principle be taken into account by weighting the expert assessments. It is e.g. possible to give more weight to the judgements of experts who are considered as more knowledgeable in the field of interest.

Also weighting may be based on the experts' performance on "seed variables". In this case, the basis for weighting is not only the substance expertise but also the ability to express the uncertainty with probabilities. As far as the generation of seed variables is concerned, there are no effective procedures on how to generate them. Cooke and Goossens (2000) suggest to use *domain seed variables* (in the expert's field of expertise) and *adjacent variables* (for which expert can give an educated guess). Cooke and Goossens (2000) recommend avoiding general knowledge variables as weighting factors in expert judgement context. Experiments indicate that experts are not better than general public on general knowledge items.

A set of assessments can be evaluated by scoring rules that are a function of the difference between the actual outcomes and the assessed probability distributions. Scoring rules may be used as a basis for rewarding assessors in order to motivate them. One example of scoring rule for discrete case is the Brier Score (Granger Morgan and Henrion 1990). Brier score separates components of assessor's performance in:

- Assessor's knowledge in the task domain.
- Measure of calibration: degree to which assessed probabilities match empirical frequency.
- Resolution: power to discriminate between different levels of probabilities.

Even if scoring rules are not directly effective in promoting accurate reporting, they are nevertheless important for evaluating the performance of assessors.

The use of empirical control assessment has been a distinctive feature of the expert elicitation methods developed and used by the Technical University of Delft, also known as the Classical Model. The performance based weighting is based on experts' assessment of "seed variables". For the weighting, two

quantitative measures of performance are used: *calibration* and *information*. *Calibration*, as mentioned before, measures the statistical likelihood that actual results correspond in a statistical sense with the experts' assessments. *Information* represents the degree to which an expert's distribution is concentrated, relative to some user-selected background measure. "Good expertise" corresponds to good calibration and high information.

The Classical Model of Goossens and Harper (1998) contains three different weighting schemes. The *equal weighting* aggregation scheme assigns equal weights to each expert. In *global* and *item weighting*, the weights are developed based on expert's performance on seed variables (see section 5.4). The performance of experts on the seed variables is assumed to reflect the performance on the variables of interest in the study. *Global* weights are determined per expert, i.e. each expert gets one weight reflecting expert's overall goodness. In item weighting, weights are determined not only per expert but also per variable. Global and item weighting techniques are called performance-based weighting techniques.

Saaty (1988) has developed a technique to compare paired data within his Analytical Hierarchy Process (AHP), which constitutes a general theory of measure. This technique was originally implemented in multi-attribute decision problems under uncertainty to establish hierarchies and sort attributes according to their importance in the decision maker's opinion. This technique has been transposed into the area of expert judgment to assign weights to different experts and has been used in many applications, mainly in the USA.

A general remark about weighting of experts' opinions is that, though the use of different weights for different experts is completely licit, many authors advice the use of equal weights. Using different weights could provide undesired problems like lack of support of some experts to the final aggregated result due to the low weight assigned to their opinions.

6.3 Aggregation of judgements

In this section, we limit our summary to objective mathematical aggregation methods. There are also other approaches, such as behavioural or consensus methods, but e.g. Mosleh et al. (1988) list several references providing evidence that mathematical aggregation methods generally yield better results. Interactive group decisions have often shortcomings, like less confident members tend to limit their participation or dominant personalities have a strong influence on others. Advantages of the mathematical (or analytical) combination are that they are easy to use, it is easy to do extensive sensitivity analyses and individual experts have no influence on the judgements of other experts after the elicitation (Bonano et al. 1990). Cooke (1991) provides a good review of various models for combining expert opinion.

The most common aggregation of judgements is averaging. The principal averaging techniques are arithmetic and geometric averaging, also known as the *linear pool* and the *log-linear pool.* In the linear pool, if the distribution elicited from an expert is f_i and its weight is ω_i then, the linear combination of all the experts' opinions is

$$f = \sum_{i=1}^{N} \omega_i f_i$$

This is normally known in probability theory as a mixture of several distributions. The log-linear pool is as follows

$$\log(f) = \sum_{i=1}^{N} \omega_i \log(f_i)$$

Note that the weights ω_i should sum up to 1. If the experts have been asked to express their uncertainty on the elicited variable in fractiles of the distribution, the full individual distributions are first formed e.g. by interpolating (linear or cubic splines, or theoretical distribution fitting most likely). If theoretical distributions are firstly fitted to the estimates of the experts, aggregation is straightforward (analytical), otherwise problems with extreme fractiles like 99% could arise, since lack of overlap in those regions could also force extrapolation, which again should be done following some well known method agreed with experts.

The experts may also be asked to provide, instead of fractiles, the parameter(s) of a distribution having the same functional form. Pulkkinen (1993) discusses aggregation schemes of such situations based on information-theoretic considerations, and provides parameters of aggregated distributions for several types of distributions.

Bayesian aggregation approaches have been summarised e.g. in Cooke (1991). Bayesian modelling requires the definition of a prior probability distribution, generated by the analysts based on available data and generalist's opinion or on previous studies, which is then updated with the expert judgements using the Bayes' theorem. The expert assessments are treated as "observations". There are several models for combining expert judgements in a Bayesian framework. One of the pioneering and most relevant models was developed by Clemen and Winkler (1985) and was further developed and interpreted by Lindley (1988).

Mosleh and Apostolakis (1988) reinterpreted Clemen and Winkler's model as if expert's estimates were a sum of the true value and an additive, normally distributed error term (the additive error model) or as the product of the true value and a multiplicative error term (the multiplicative error model).

Within the Benchmark Exercise on Expert Judgement Techniques (Cojazzi and Fogli 2000) two partners, UPM and STUK, used a Bayesian approach for aggregation. The basic principle in the VTT-STUK approach was that the variable

of interest is assumed to be a random variable with some distribution with unknown parameters. It was assumed that the quantiles elicited from experts represent a sample of quantiles of a finite but unknown sample from the same distribution. Since the parameters of the distribution are unknown, they are modelled wit suitable distributions. Here non-informative prior distributions were applied. These distributions are updated with the expert judgements to obtain posterior distributions of the parameters, which then update the uncertainty distribution of the variables of interest.

A final comment should be made about the issue of deep disagreement among experts. In some applications, even after some reconciliation sessions, huge disagreements between experts do remain (non-overlapping or poorly overlapped distributions). In those cases a purely mathematical aggregation could be risky, since the result of some aggregation techniques could lead to a result in which values considered as ‘impossible’ or ‘almost impossible’ by all experts could get some likelihood. Those situations, in our opinion, should be avoided. A possible option is to keep the different distributions apart and to use them for sensitivity studies.

7 APPLICATIONS OF EXPERT JUDGEMENT

Expert opinion has been used in more or less formal way in many fields. Important applications have been in the aerospace field, in military intelligence, in nuclear energy and in policy analysis.

7.1 Non nuclear applications

The use of expert opinion in the aerospace sector was linked to the need to assess risks associated with rare or unobserved catastrophic events. The problems of assessing likelihood of such events, which is not possible through the traditional scientific method of repeated independent experiments, were dramatically brought to attention after the Challenger space shuttle accident in 1986.

In the Netherlands, expert opinion has been used in different fields. In particular, the University of Delft have been in the forefront in developing and applying expert judgement. Some examples of applications in different fields are given below (Cooke and Goossens, 2004):

- In the chemical field, expert judgement was applied in the ranking of management factors and relative failure frequencies (such as processes, storage and transport) in chemical installations falling under the EU Seveso Directive. Also it was applied to the assessment of hazardous airborne particles under various meteorological circumstances.
- In the aerospace sector expert judgement exercises were done to determine predictions for space debris loads and space shuttle composite material.
- In the veterinary sector, expert judgement was used to determine model parameters for respiratory diseases. The quantitative experts' assessments were used as input in economic models for farming practices.
- Some expert judgement exercises have also been performed for supporting reliability of critical infrastructures. For example, assessment of dike-ring failures probability in the Netherlands in order to prevent flooding; assessment of the reliability of movable water barriers and assessment of contributing factors to accident frequency in inland waterway transportation by ships.

Expert judgement is also widely used in policy analysis. Some examples reported in Granger Morgan and Henrion (1990) refer to studies on the assessment of lead exposure and on the depletion of stratospheric ozone by chlorofluorocarbons. However, the field of policy analysis is extremely wide and contains many methodologies that not always can be called "expert opinion" analysis.

7.2 Nuclear applications

The first application of subjective probabilities in probabilistic risk analysis took place in the Reactor Safety Study WASH 1400 published by the American Nuclear Regulatory Commission in 1975 (WASH 1400, 1975). This study was considered to be the first modern Probabilistic Risk Assessment and made large use of subjective probabilities.

After that, expert opinion has been used in a structured form as a source of data in several large studies in the nuclear field. Among these we can name studies on seismic risk, on fire hazards in nuclear power plants in the USA and on risks assessment. The Reactor Risk Reference Document (NUREG-1150, 1990) uses extensively expert judgement to assess risks of five nuclear power plants.

In the area of Radioactive Waste Repositories, many expert judgement applications have been conducted, as well in USA as in Europe. In USA, regarding the Hanford site (DOE/RW-0017, 1984, and Golder associates, 1986), in addition to utility functions, expert judgment was used to elicit distributions for parameters regarding the behaviour of groundwater flows and gas solutions in water. DOE (DOE/RQ-0074, 1986) did also perform an expert judgement application to classify five potential sites for the High Level Waste (HLW) repository for the civil nuclear programme. Also the report NUREG/CR-5411 (Bonano and al. 1990) describes a study on the use of expert judgement in performance assessment for high-level radioactive waste repositories.

Extensive expert judgement studies have also been applied to the Waste Isolation Pilot Plant (WIPP) in 1979 by Sandia National Laboratory. The Performance Assessment process had to demonstrate that WIPP complies with long-term environmental standards. It used mathematical models to predict consequences of various scenarios over a period of 10000 years. Expert judgement has been used for the design of such relevant scenarios. An expert elicitation procedure was also applied in eliciting information about marking the WIPP to deter human intrusion (Rechard and al., 1993). Expert judgement applications have been also applied to the assessment of parameters related to solubility and sorption of different radionuclides.

Several American institutions, like Rockwell International, SNL and Southwest Research Institute (SRI), have performed expert judgement studies related to the Yucca mountain site. Expert judgement was used in the Basalt Waste isolation Project (BWIP) to elicit the porosity of the medium and the anisotropy coefficient. SNL performed successive Total System Performance assessments in 1991 and 1993. In the first study hydrology parameters, such as sorption coefficients and percolation rates were assessed. In the second one, sorption coefficients and solubility coefficients for different radionuclides were assessed. Additionally, SRI got also some estimates through expert judgement of possible climatic conditions for the next 10000 years in this site, and EPRI used the same kind of techniques

to assess the probability of earthquakes for the same site during the same period.

The Department of Environment (DOE) of the UK collaborated with other European institutions in the EU project PACOMA, proposing expert judgement techniques as adequate tools to address the issue of uncertainties in the Performance assessment of HLW repositories. These techniques were applied to the case of Harwell site. This approach was further used by her Majesty's Inspectorate of Pollution (HMIP) in the Dry Run 3 study. While in PACOMA the focus was on geospheric parameters like dispersivities, diffusivities and hydraulic conductivities, in Dry Run 3 the expert judgement techniques were focussed on environmental and biosphere changes.

Expert elicitation techniques were benchmarked in the "Benchmark exercise on Expert Judgement Techniques in PSA Level 2". During the benchmark two different case studies were submitted to experts: (1) assessment of phenomenology of fuel-coolant interaction in the case of a severe accident and (2) assessment of hydrogen deflagration/detonation due to zirconium cladding oxidation during core reflood following power recovery after a station black-out initiating event.

7.3 Applications in the field of NPPs structural reliability

Estimation of failure probabilities in PWR pressure boundary system components

The Pacific Northwest Laboratory (PNL) conducted in 1990 an expert judgement elicitation to obtain numerical estimates for probabilities of catastrophic or disruptive failures in pressure boundary systems and components in PWRs. The results were presented in a paper (Vo et al. 1991), which is summarised in the following.

Scope of the study

For the analysis, following systems from the Surry Unit 1 were selected:

- Reactor pressure vessel.
- Reactor coolant system.
- Low pressure injection system including the accumulators.
- Auxiliary feed water systems.

For the reactor pressure vessel, components of interest were the vessel shell, heads, flanges, closure studs, penetrations, nozzles, safe ends and attachment welds.

For the other systems, the components of interest were piping segments. These might include the straight lengths of pipe, pipe elbows, couplings, fittings, flanged joints and welds. Failures in piping of less than one-inch diameter were excluded from the study, as well as active components such as pumps and valves.

Only structural failures perceived as important to plant risk, or that could significantly affect the core damage frequencies, were considered. The primary concern was a sudden and complete severance (rupture) of components or structures in the PWR systems. Component ruptures are event involving severe leakage or catastrophic failure of pressure boundary structures or components that could disable the intended function the safety systems.

Expert elicitation

The expert elicitation was performed using a systematic procedure, which closely followed the approach used in the NUREG-1150. The selected experts needed to have knowledge of the subject matter, and they had demonstrated their expertise by publications, hands-on experience and managing or performing research in the areas related to the issue. A balanced representation of experts from academia, private industry, and government was involved in the study.

The study consisted of a three-day elicitation meeting. At the beginning, objectives were outlined and the types of information sought were described. As background information, the risk-based methodology and the elicitation process were also described. However, no formal training was provided to the experts. Most of the time was devoted to eliciting issues and discussions.

For each issue, a formal presentation was provided, including technical descriptions, historical component failure mechanism, elicitation statements, suggested approaches, questionnaire forms and additional material. After the formal presentations and discussions on issues, the elicitation statements were presented. The following is an example of elicitation statement:

> *"On the attached questionnaire form, assess the rupture probability (or rupture rate) for each piping segment and its associated uncertainties using data and information discussed during the elicitation meeting.*
>
> *In all cases, please provide the rationale of your assessment"*

The information used to support the experts for estimating the desired inputs was the following:

- Historical failure data
- Data from fracture mechanics analyses
- PSA results and other relevant information (system, component prioritisation, system descriptions etc.)

- Additional plant specific information, etc.

Experts were asked to fill out specific questionnaire forms. For each piping segment or component, experts provided three estimates: a best estimate and uncertainty bounds. It is not explained in the paper how these uncertainty bounds were interpreted.

After the elicitation meeting, the information provided by the expert panel was recomposed and aggregated. The expert panels reviewed the written analyses of each issue. As the experts from one plant could not attend the elicitation, additional plant-specific data obtained in a post-elicitation meeting was included afterwards, and the panel was asked to revise their original estimates, if needed. The revised individual judgements were again aggregated.

Aggregation of judgements and results of the study

Following the revision of the expert judgements, PNL staff compiled the results into statistical distributions. In the aggregation of expert judgements, the median or geometric mean was used. The population quartile was chosen to describe the uncertainty in the estimate.

After the aggregation of judgements, the “best estimates” obtained from experts were summarised and presented graphically with box and whiskers plots. The whiskers identify extreme upper and lower values and the box locates the lower and upper quartiles. Also the median of the distribution is indicated.

As a summary, the results appeared to be reasonable, estimates were wide spread and generally agreed with and reflected plant operating experience.

Conclusions regarding the expert judgement procedure

The paper concludes that the experts were excellent and their opinions appeared to be unbiased and knowledgeable. It was recommended that more early attention should be given to the structure of questions. No formal training was provided to experts, which is clearly a shortage. The paper suggests that at least a half day of formal training should be included in future elicitations.

Development of passive system LOCA frequencies using expert elicitation

A very recent expert judgement study related to passive components of nuclear power plants is the USNRC project on Loss of Coolant Accident (LOCA) frequency development (USNRC 2004). The objective of the project was to develop piping and non-piping passive system LOCA frequency distributions as a function of rupture size and operating time from the current day up to the end of the license extension period.

The project recognises that an expert judgement approach is a natural approach for development of the LOCA frequencies due to the scarcity of available data and the subject complexity. Traditionally LOCA frequencies have been assessed either by statistically analysing service experience data or by performing probabilistic fracture mechanics analyses of specific postulated failure mechanisms. Both of these approaches have limitations, and the expert elicitation is aimed at mitigating deficiencies in them.

The formal judgement approach in this study consisted of the following steps:

- Conduct pilot elicitation.
- Select panel and facilitation team.
- Develop technical issues:
    - Construct approach for estimating LOCA frequencies.
    - Determine significant issues affecting LOCA frequencies.
- Quantify base case frequencies:
    - Develop estimates for well-defined piping conditions.
    - Two estimates used probabilistic fracture mechanics analysis and two estimates used operating experience analysis.
- Formulate elicitation questions.
- Conduct individual elicitations.
- Analyse quantitative results and qualitative rationale.
- Summarize and document results.

The first step in the study was to conduct a pilot elicitation using NRC staff members. This pilot exercise identified important technical issues for consideration and provided feedback to design the approach for the formal elicitation.

For the elicitation panel (expert panel), 12 experts were selected among 55 nominally qualified candidates. The potential panel members were affiliated within industry, academia, national laboratories, contracting agencies, and other government and international agencies. The final panel members were chosen to represent a range of relevant technical specialities, and they all had at least 25 years of experience in the relevant technical areas.

In addition to the expert panel, a facilitation team was assembled. This team, consisting of one normative expert, six substantive experts and two recorders, guided the expert panel through the elicitation process. The team formulated the objectives and scope, provided background technical information, developed elicitation questions, guided and recorded the individual elicitation sessions, analysed and summarised the panel's findings, and developed the final LOCA frequency distributions from the panel's responses.

A three-day kick-off meeting of the expert panel and the facilitation team was held in February 2003, and it had the following principal objectives:

- Definition of the scope and objectives of the expert judgement process.
- Providing background information about previous LOCA frequency estimates.
- Construction of an approach for determining LOCA frequencies.
- Identification of significant issues affecting LOCA frequencies.
- Conducting elicitation training.

The flowchart in figure 4 presents the analysis of the elicitation responses.

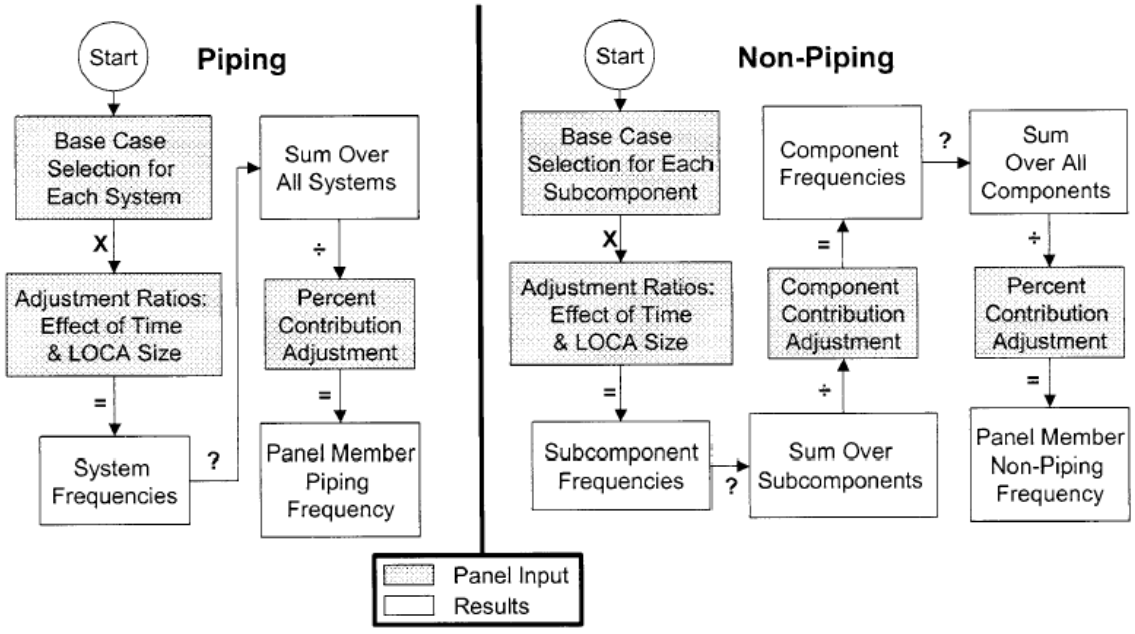


Figure 4: Analysis of elicitation responses: flowchart
NRC Development of Passive System LOCA Frequencies for Risk-Informed revision of 10 CFR 50.46 (R. Tregoning and L. Abramson, 2004)

The panel defined six LOCA categories to be evaluated. Large break LOCA regime was divided into four categories to reflect the different plant responses for mitigating LB LOCA events. The panel developed a structure for considering passive system failures that contribute to LOCAs. The total passive system frequencies were divided into piping and non-piping contributions.

Four panel members were selected to develop base case frequencies to provide the panel with quantitative estimates for anchoring their responses. Five base cases were defined for three PWR piping and two BWR piping systems.

The various approaches used to estimate these frequencies were presented in a second meeting by the base case developers. Each panel member chose a base case and developed individual estimates by anchoring their evaluations on the base cases.

The panel members were elicited individually. The objectives of the elicitation session were the following:

- Obtain and discuss the qualitative and quantitative responses to elicitation questions.
- Identify inconsistencies between the qualitative and quantitative responses.
- Provide additional clarification to the elicitation questions, as necessary.
- Identify necessary follow-on work for each panel member.
- Solicit feedback about the process

In the elicitation, the experts were asked to provide the median and the 5^{th} and 95^{th} percentiles for each question. During these sessions, weaknesses, inconsistencies or incomplete areas were identified for each panel member. The experts had then one to four months time to revise the initial input to address any deficiencies.

The percentiles provided by the panel members were treated as percentiles of log-normal distributions. Split distributions were used to accommodate asymmetric responses. These distributions were combined to obtain aggregated distributions.

The third meeting with the entire panel was held when individual elicitations had been completed and the results had been analysed. The purpose of the meeting was to summarise the qualitative and quantitative results form elicitations, and to present the methodology for aggregation of judgements. The panel members were given the opportunity to revise the estimates based on qualitative rationale from other experts.

Other studies

In the mid-90s an expert judgement process was developed and applied for structural failure rates of components in the System 80+ advanced reactor design (Bullough et al. 1999). In this process, the expert elicitation was done by paired comparison.

DISCUSSION AND CONCLUSION

This document aims at providing an overview on formal expert judgement, and it is targeted to readers that are not very familiar with the issue, but are interested to have a condensed summary on the topic. An emphasis has been put on the use of formal expert judgement in the field of structural integrity, this being an area of interest in plant life management of ageing nuclear power installations.

In the field of structural integrity there is a need to move towards more realistic estimates where uncertainties are modelled with probabilistic approaches. One example of such need for probabilistic evaluation is the risk-informed in-service inspection (RI-ISI) methodology, where the aim is to optimise inspections on the base of risk importance of piping (or other structural components). In order to evaluate the risk, estimates of the probability of failure and of the failure consequences are needed. There are several degradation mechanisms for which no validated structural reliability tools are available. Also, even for better-known mechanisms, different models may produce quite different results. The scarcity of operating experience and the quality of data may limit the use of statistical approaches. In order to obtain numerical estimates for structural reliability issues, the use of structured expert judgement may be the best way forward.

Although formal expert judgement has become a rather well established tool in connection to risk assessments, and the main steps in such expert judgement processes are quite similar, there is variability related to some important issues. We discuss below some of such issues, and present our position that will be used at least in our first case study.

One question is how much guidance or structure should the experts be given for solving the problem. One extreme is that the model or decomposition of the problem has been fixed, and the experts only assess the uncertainty distributions of some variables. The other extreme is to give the experts free hands to decide on the approach how to solve the problem. We tend to favour the latter option, leaving more room for different approaches to be presented. However the choice of the most suitable approach is dependent on the problem and should be decided case by case.

A controversial issue is the scoring or weighting of experts since fairness and adequacy of the calibration variables can easily be questioned. For the time being, we will restrain from using weighting of experts based on exercises. Weighting can however be used to study the sensitivity of the aggregated results on various expert judgements.

There are different opinions on the manner to conduct the elicitation of experts, and whether the experts should interact with each other or not. We believe that it is useful to have a meeting with all the experts to discuss the approaches they have used to solve the problem. However, the experts should not discuss their

numerical judgements in this connection. The elicitation of final results should be done individually.

The aggregation of judgements can be done using several approaches. We recognise that different aggregation techniques may lead to quite different aggregated distributions, and thus this is an issue of scientific debate. However, we are not going to express any strong opinion on this issue at this phase.

A formal expert judgement process is led by at least one normative expert. In the case of USNRC project on LOCA frequency development, the so-called “facilitator team” consisted of one normative expert, six substantive experts and two recorders. The composition is naturally dependent on the resources available, but should also reflect the complexity of the issues to be analysed. In the case of structural reliability, it would be recommended that the expert judgement process be led by a team consisting of at least one expert in this field (substantive expert) working together with the normative expert(s).

This document has an emphasis on structural integrity, however it can also be seen as a basis for developing an approach for formal expert judgement in other plant life management areas such as, for example, maintenance and in-service inspection.

REFERENCES

Bolado, R., and Gallego, E. (2000). *El Juicio de Expertos y su aplicación a cuestiones de seguridad.* Fundación para el Fomento de la Innovación Industrial.

Bonano, E.J., Hora, S.C., Keeney, R.L., and von Winterfeldt, D. (1990). Elicitation and Use of Expert Judgement in Performance Assessment for High-Level Radioactive Waste Repositories. *U.S. Nuclear Regulartory Commission, NUREG/CR-5411.*

Brown, B., Cochran, S., and Dalkey, N. (1969). *The Delphi Method II: Structure of Experiments*. The RAND Corporation

Bullough R., Green V.R, Tomkins B., Wilson R., Wintle J.B. (1999). A review of methods and applications of reliability analysis for structural integrity assessment of UK nuclear plant. *International Journal of Pressure Vessels and Piping,* Vol. 76, 909-919.

Clemen, R.T., and Winkler, R.L. (1985). Limits for the Precision and Value of Information from Dependent Sources. *Operations Research,* Vol. 33, 427-442.

Cojazzi, G., and Fogli, D. (2000). Benchmark Exercise on Expert Judgement Techniques in PSA Level 2, Extended Final Report. *EUR Report 19739 EN*

Cooke, R. M., (1991). Experts in uncertainty. Opinion and Subjective Probability in Science. Oxford University press.

Cooke, R. M., and Goossens, L.H.J. (2000). *Procedures guide for structured expert judgement*. EUR 18820 EN

Cooke, R. M., and Goossens, L.H.J. (2004). Expert judgement elicitation for risk assessment of critical infrastructures. *Journal of Risk Research* 7 (6), 643-656 September 2004

Dalkey, N.C. (1969). *The Delphi Method: An experimental study of group opinion*, The RAND Corporation

DOE/RQ-0074 (1986). *A Multiatribute Utility Analysis of Sites Nominated for Characterization for the First radioactive Waste Repository*. A Decision aiding Methodology. U.S. Department of energy.

DOE/RW-0017 (1984). Draft *Environmental Assessment. Reference Repository Location. Hanford Site.* U.S. Department of Energy.

Golder Associates (1986). *Report to Rockwell Hanford Operations – Probability Encoding of Variables Affecting Inflow into the Exploratory Shaft Test Facility at the Basalt Waste Isolation Project.* Ridgeland, Washington.

Goossens, L.H.J. and Harper, F.T. (1998). Joint EC/USNRC expert judgement driven radiological protection uncertainty analysis. *Journal of Radiological Protection*, Vol.18, No. 4, 249-264.

Gordon, T. J. (1994). *Cross Impact Method*. Futures Research Methodologies. AC/UNU Millennium Project

Granger Morgan, G. M., and Henrion, M. (1990). *Uncertainty. A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analy*sis. Cambridge University Press

Hampton, J.M., Moore, P.G., and Thomas H. (1973). Subjective Probability and its Measurement. J. R. Statist. Soc. A, 136, Part 1, 1973, 21-42.

Hogarth, R.M. (1975). *Cognitive Processes and the Assessment of Subjective Probability distributions*. JASA, Vol. 70, No. 350, 271-294.

Kahn, H. (1960). *On Thermonuclear War*, Free Press, New York

Kahneman, D., Slovic P. and Tversky, A. (1982). *Judgement under uncertainty: Heuristics and biases*. Cambridge University press.

Merkhofer, M. W., (1987). Quantifying Judgement Uncertainty: Methodology, Experiences, and Insights. *IEEE Transactions on Systems, Man, and cybernetics.* Vol. SMC-17, No. 5

Lindley, D.V. (1988). The Use of Probability Statements. In Proceedings of the CII Course of the International School of Physics 'Enrico Fermi' on *Accelerated Life Testing and Expert Opinions in reliability*.

Mosleh, A., Bier, V. M., and Apostolakis G. (1988). A Critique of Current Practice for the Use of Expert Opinions in Probabilistic Risk Assessment. *Reliability Engineering and System Safety 20 (1988) 63-85*

Mullin, T.M. (1986). *Understanding and Supporting the Process of Probability Estimation*. PhD. diss. Carnegie Mellon University, Pittsburgh.

NUREG-1150, U.S.Nuclear Regulatory Commission, December 1990. *Severe Accident Risks: An Assessment for Five U.S. Nuclear Power Plants*.

Otway, H. and von Winterfeldt, D. (1992). Expert judgement in risk analysis and management: process, context and pitfalls. *Risk Analysis*, Vol.12, No 1.

Pulkkinen, U. (1993). Statistical Models for Expert Judgement and Wear Prediction. *VTT Publications 181. Technical Research Centre of Finland.*

Rechard, R. P., Trauth, K. M., Rath, J. S., Guzowski, R. V., Hora, S. C., and Tierney, M. S. (1993). The Use of Formal and Informal Expert Judgement When Interpreting Data for Performance Assessments. *SAND92 – 1148* .

Saaty, T.L. (1988). Mathematical methods of Operations Research. Ed. Dover.

Slovic, P., Finucane, M.L., Peters, E., and MacGregor, D. G. (2004). Risk as Analysis and Risk as Feelings: Some Thoughts about Affect, Reason, Risk, and Rationality. *Risk Analysis*, Vol. 24, No. 2

Slovic, P., and Fischoff, B., (1977). On the Psychology of Experimental Surprises. *Journal of Experimental Psychology: Human Perception and Performance*, 3 no. 4, 544-551

Tregoning, R. and Abramson, L., (2004). *Development of Passive System LOCA Frequencies for Risk Informed Revision of 10 CFR 50.46*. ACRS Subcommittee on Regulatory Policies and Practices, 1st April 2004

USNRC (2004). U.S. Nuclear Regulatory Commission 10 CFR 50.46 LOCA frequency development. Attachment.

Vo, T.V., Heasier, P.G., Doctor, S.R., Simonen, F.A. and Gore, B.F. (1991). Estimates of component rupture probabilities: expert judgement elicitation. PVP-Vol. 215, *Fatigue, Fracture, and Risk, ASME*.

WASH 1400, U.S. Nuclear Regulatory Commission, 1975. *Reactor Safety Study: An Assessment of Accident Risks in U.S. Commercial Nuclear Power Plants*.

Woo, G., Chamley, J. E. and Taylor, J. (1992). The use of Expert Judgements in Risk Assessment. DOE Report No. DoE/HMIP/RR/92.058.

Winkler, R.L., (1967). The Assessment of Prior Distributions in Bayesian Analysis. JASA. September 1967, 776-800.

European Commission

EUR 21772 EN – DG JRC – Institute for Energy

Formal Expert Judgement
An Overview

Authors:

K. Simola
A. Mengolini
R. Bolado

Abstract

This document provides an overview on the formal process of expert judgement for readers who are not familiar with the issue, but wish to have a condensed summary of the topic. An emphasis is put on the use of formal expert judgement in the field of structural integrity, an area of interest in plant life management of ageing nuclear power installations.

The mission of the Joint Research Centre is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of commercial or national interests.